Attrition suffered at work

Statistical methods for Data Analytics

Ricardo Pinto

Contents

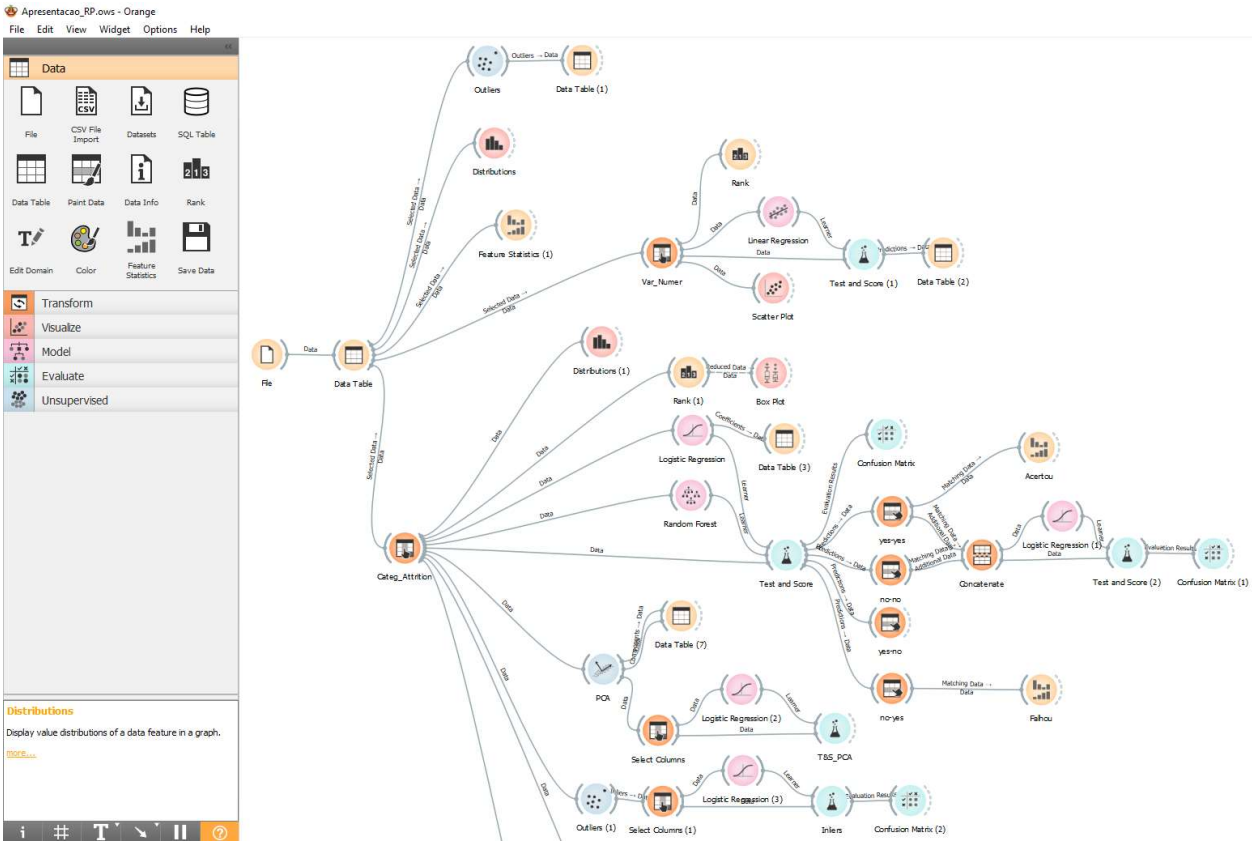
Attrition suffered at work 1
1. INTRODUCTION 2
2. DESCRIPTIVE STATISTICS 2
2.1. Distribution Histograms 2
2.2. Numerical Attributes description 5
3. HYPOTHESIS TESTS 6
3.1. Normality 6
3.2. Gender 7
4. LINEAR REGRESSION 10
4.1. Numerical Target 10
4.2. Categorical Target 11
5. REGRESSION MODELS 13
6. CLUSTERS 17
Conclusion 20

1. INTRODUCTION

This dataset reports on factors such as employees' age, gender, salary, job role and satisfaction, and asks to relate these to attrition. The purpose is therefore to study these variables and how they relate to attrition suffered at work.

It includes 1470 workers and 33 attributes. The target variable is, as mentioned above, “attrition”. It is a categorical variable (“yes”, “no”) and therefore nominal.

For statistical and data analysis, we used Orange and SPSS programs. In Orange desktop we’ve built the following scheme:

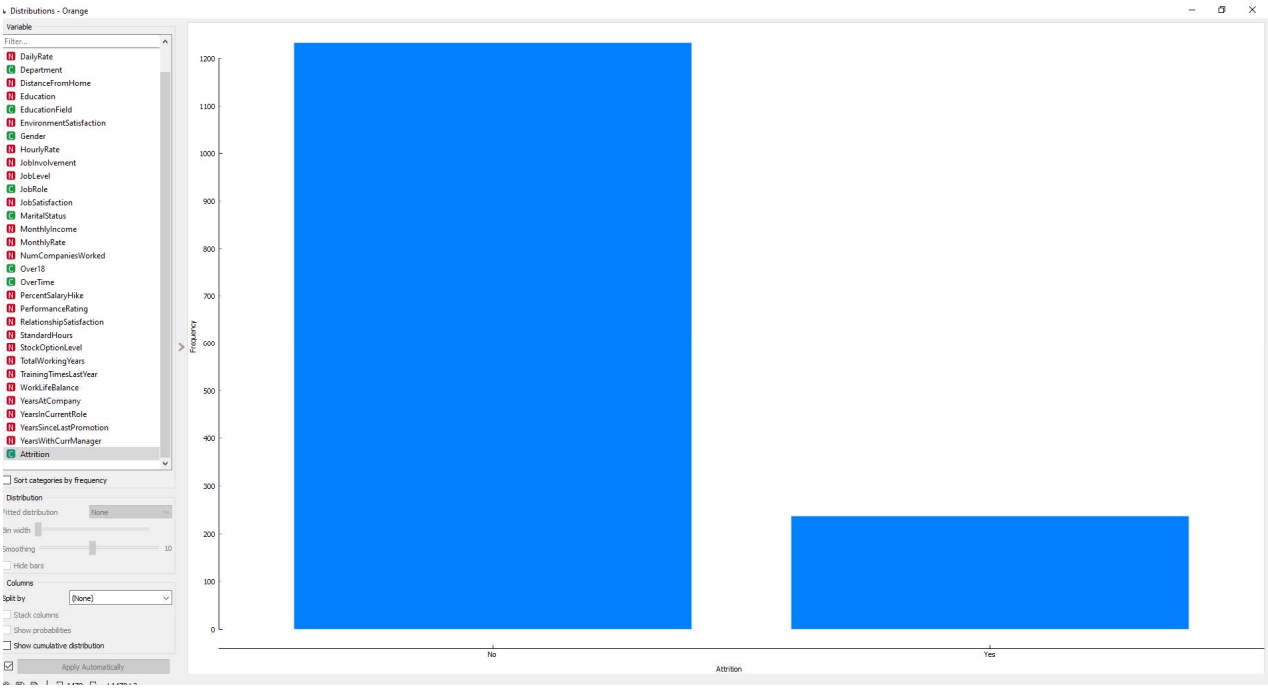


SPSS was required to carry out statistical analysis, namely specific hypothesis tests.

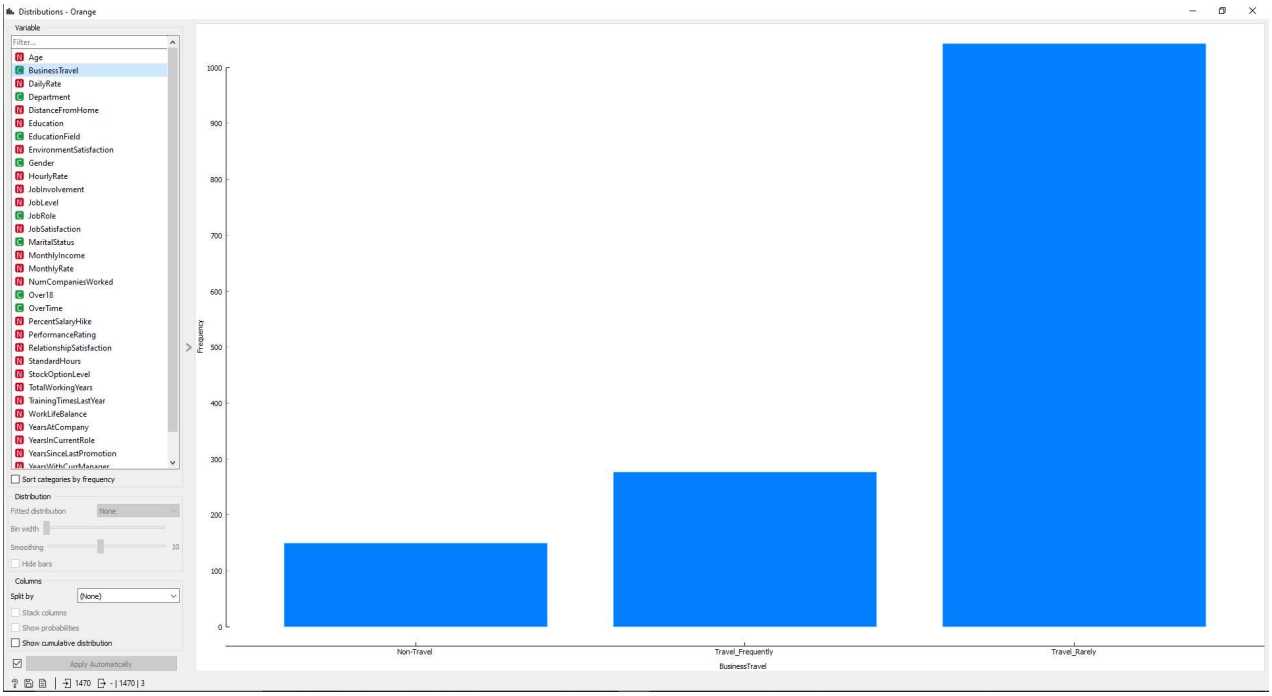
2. DESCRIPTIVE STATISTICS

2.1. Distribution Histograms

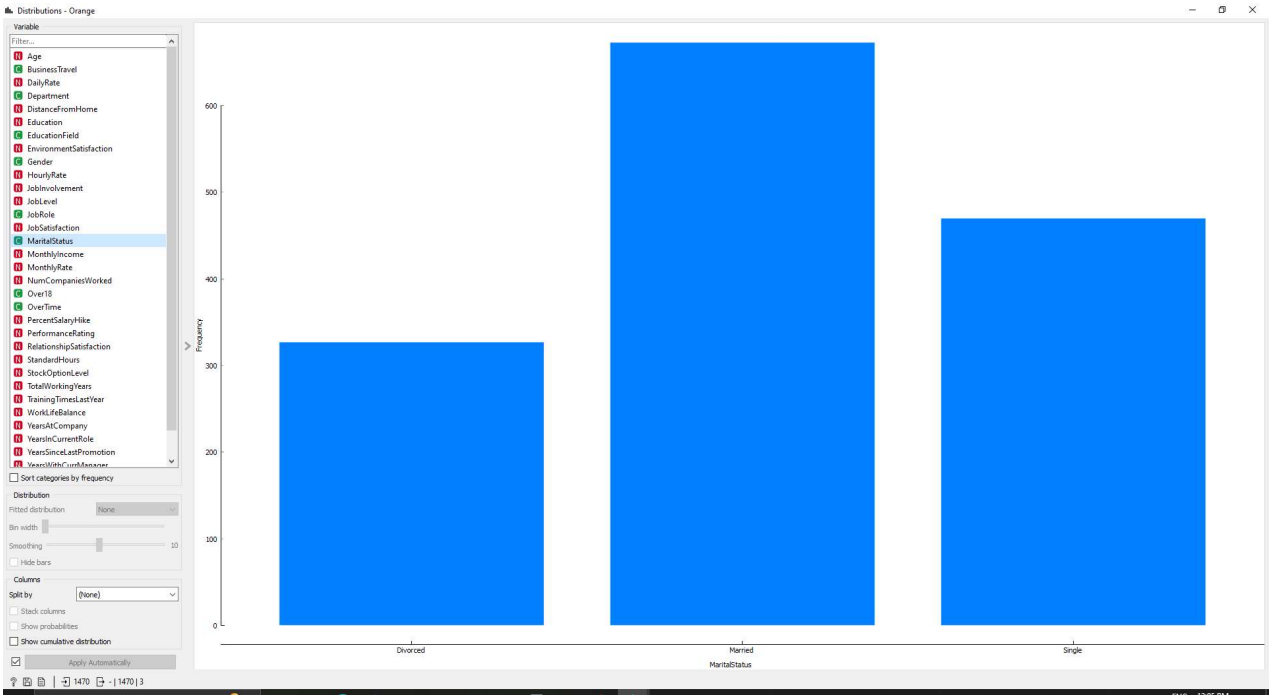
Most of the workers do not suffer attrition.



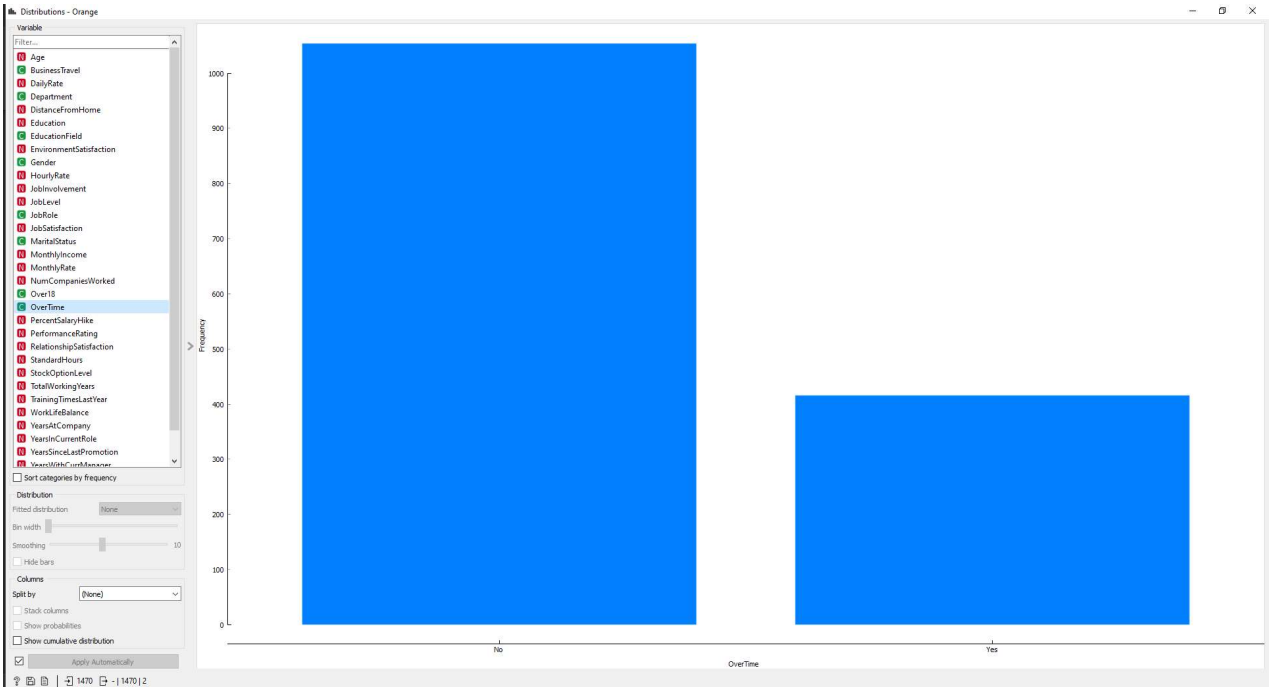
Business travel: most of the workers don’t travel



Marital status: more married than single



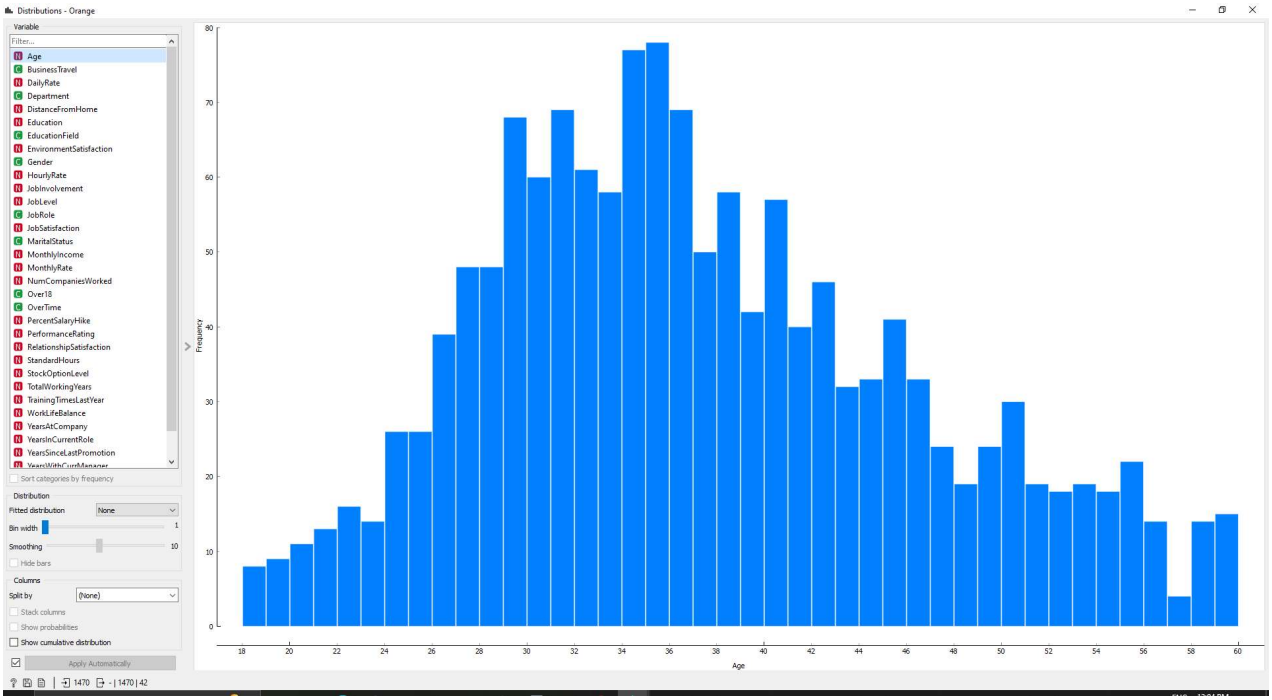
Overtime: the majority of workers do not work overtime



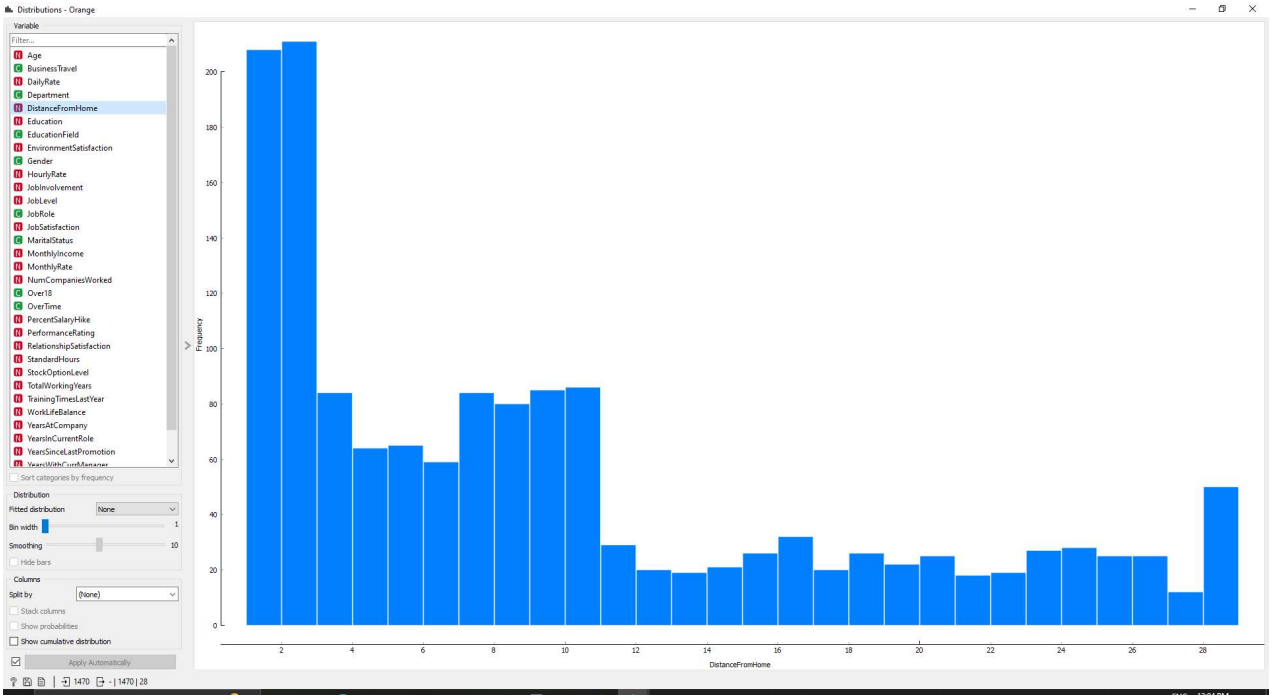
2.2. Numerical Attributes description

Age:

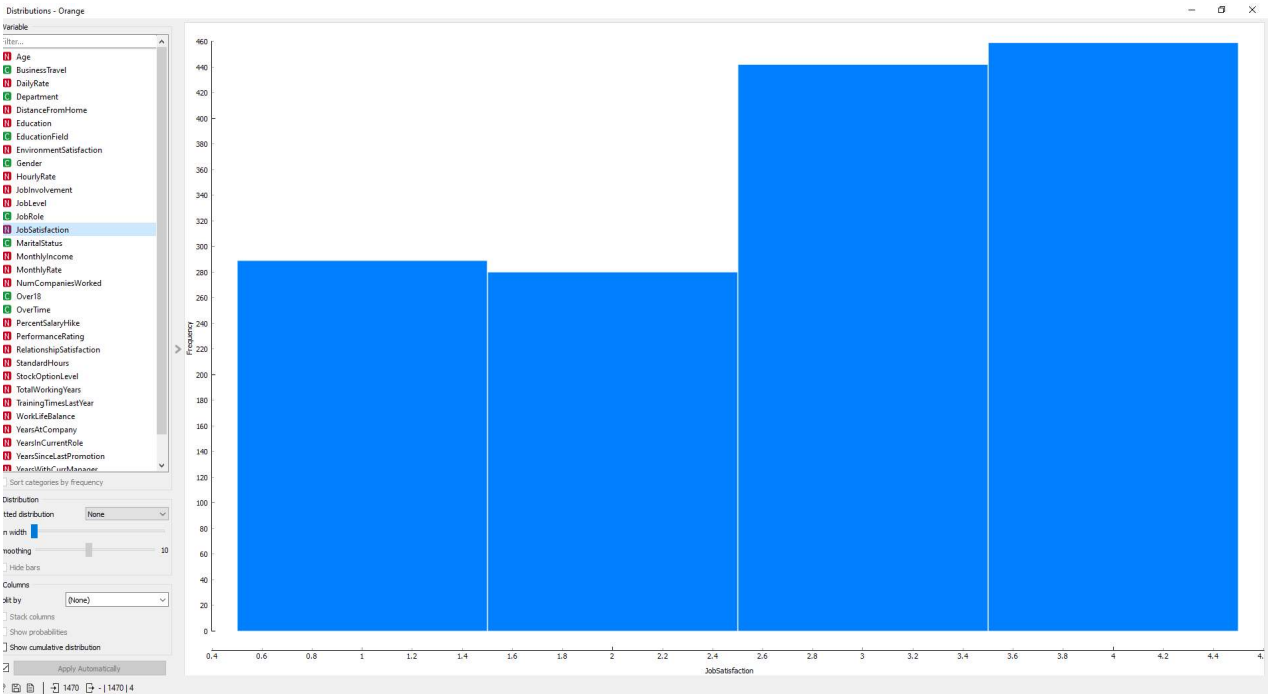
The average age is aprox. 37 years old.
Median is 36 years old.
The distribution has a slight skewed deviation to the left.



Distance from home: more people are living close to work



Job satisfaction: more people are satisfied than dissatisfied



3. HYPOTHESIS TESTS

3.1. Normality

We checked in SPSS if the distributions of the variables are normal or not.
We did 1 hypothesis test in which we started from the H0 hypothesis in which the distribution is normal, and H1 is not normal.
In all variables, it was verified in the Kolmogorov test that the significance = 0.0 or almost = 0, which means we can reject the hypothesis that the distribution is normal for all variables.

Testes de Normalidade

		Kolmogorov-Smirnov[a]			Shapiro-Wilk	
	Attrition	Estatística	gl	Sig.	Estatística	gl
Age	No	,082	1233	,000	,978	1233
	Yes	,120	237	,000	,946	237
DailyRate	No	,073	1233	,000	,954	1233
	Yes	,075	237	,003	,950	237
DistanceFromHome	No	,162	1233	,000	,853	1233
	Yes	,138	237	,000	,896	237
Education	No	,223	1233	,000	,897	1233
	Yes	,247	237	,000	,886	237
EnvironmentSatisfaction	No	,215	1233	,000	,849	1233
	Yes	,198	237	,000	,837	237
HourlyRate	No	,071	1233	,000	,954	1233
	Yes	,078	237	,001	,956	237
JobInvolvement	No	,339	1233	,000	,804	1233
	Yes	,315	237	,000	,821	237
JobLevel	No	,267	1233	,000	,835	1233
	Yes	,354	237	,000	,701	237
JobSatisfaction	No	,210	1233	,000	,843	1233
	Yes	,210	237	,000	,851	237
MonthlyIncome	No	,176	1233	,000	,834	1233
	Yes	,180	237	,000	,780	237
MonthlyRate	No	,068	1233	,000	,954	1233
	Yes	,080	237	,001	,953	237

3.2. Gender

We also did another hypothesis test with the following question: Are the male and female groups significantly different in terms of monthly income?

As male and female groups are independent, we applied the T test.

Levene's test and T test show a Sig > 0.05, so we accept that there is no difference between genders in terms of salary.

Teste de amostras independentes

		Teste de Levene para igualdade de variâncias		teste-t para Igualdade de Médias						
									95% Intervalo de Confiança da Diferença	
		Z	Sig.	t	df	Sig. (2 extremidades)	Diferença média	Erro padrão de diferença	Inferior	Superior
MonthlyIncome	Variâncias iguais assumidas	,259	,611	-1,221	1468	,222	-306,058	250,608	-797,647	185,530
	Variâncias iguais não assumidas			-1,222	1261,547	,222	-306,058	250,403	-797,311	185,194

Another alternative, given that the distributions are not normal, would be to do a nonparametric test. As they are independent samples, the most suitable one is the Mann-Whitney test.

The results confirm the hypothesis H0 should be retained, which says the monthly income is not significantly different between men and women.

Testes não paramétricos

Resumo de Teste de Hipótese

	Hipótese nula	Teste	Sig.
1	A distribuição de MonthlyIncome é igual nas categorias de Gender.	Amostras Independentes de Teste U de Mann-Whitney	,088

Resumo de Teste de Hipótese

	Decisão
1	Reter a hipótese nula.

São exibidas significâncias assintóticas. O nível de significância é ,050.

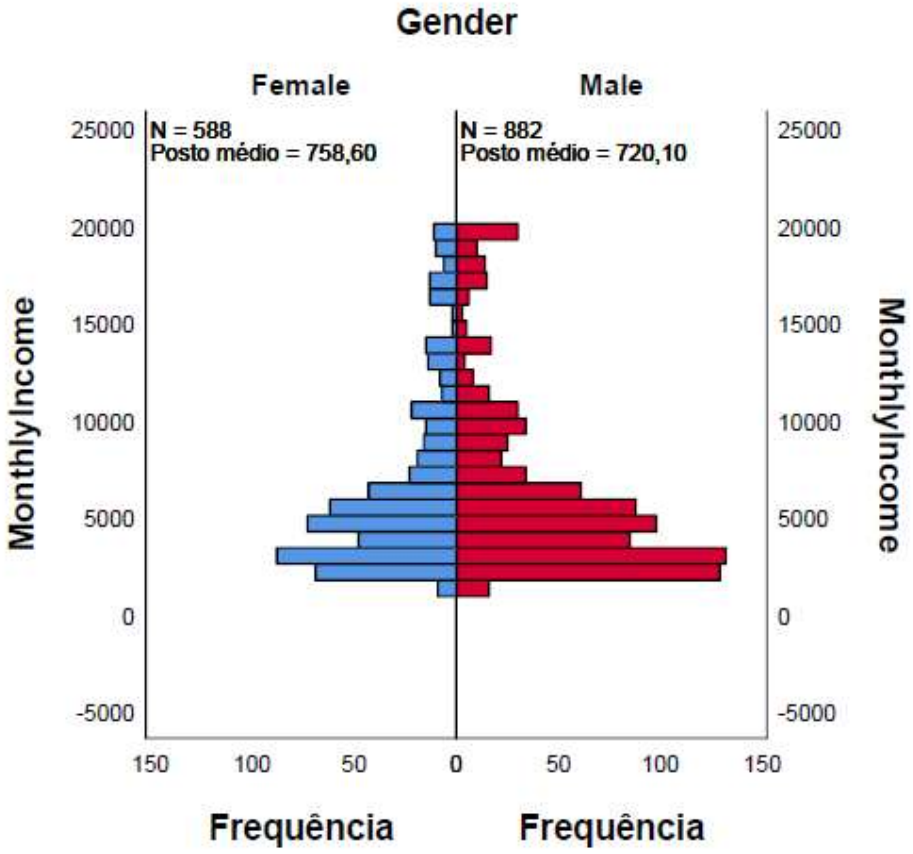
Amostras Independentes de Teste U de Mann-Whitney

MonthlyIncome entre Gender

Amostras Independentes de Resumo de Teste U de Mann-Whitney

N total	1470
U de Mann-Whitney	272893,500
Wilcoxon W	446059,500
Estatística do teste	272893,500
Erro padrão	7973,309
Estatística de Teste Padronizado	1,704
Sinal assintótico (teste de dois lados)	,088

Amostras Independentes de Teste U de Mann-Whitney

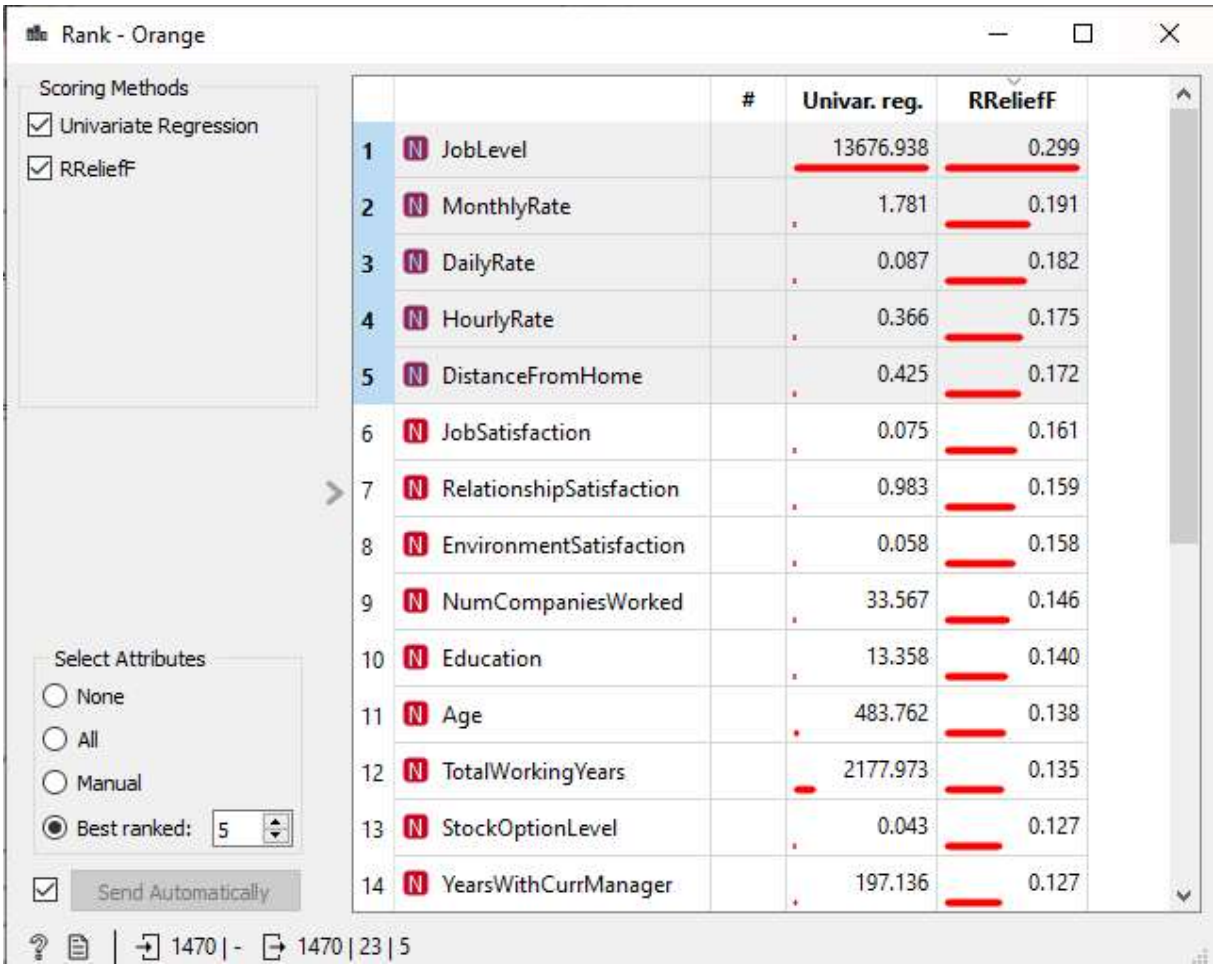


4. LINEAR REGRESSION

4.1. Numerical Target

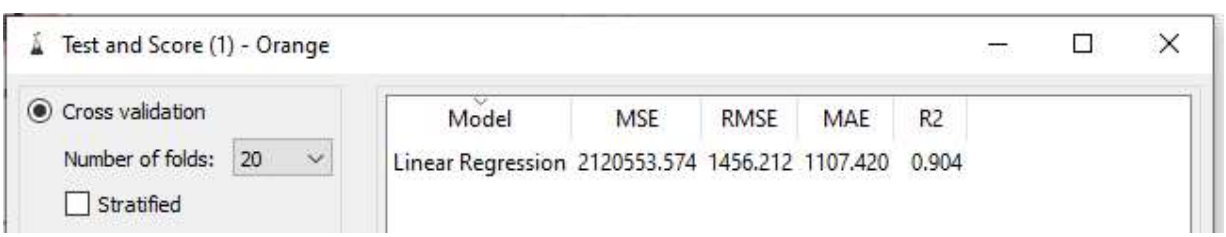
I can also analyze the monthly income based on the remaining variables, to assess which ones are most correlated with salary, through multiple linear regression.

We can see that variables such as job level, rate and distance from home are the ones that have the strongest correlation with monthly income.



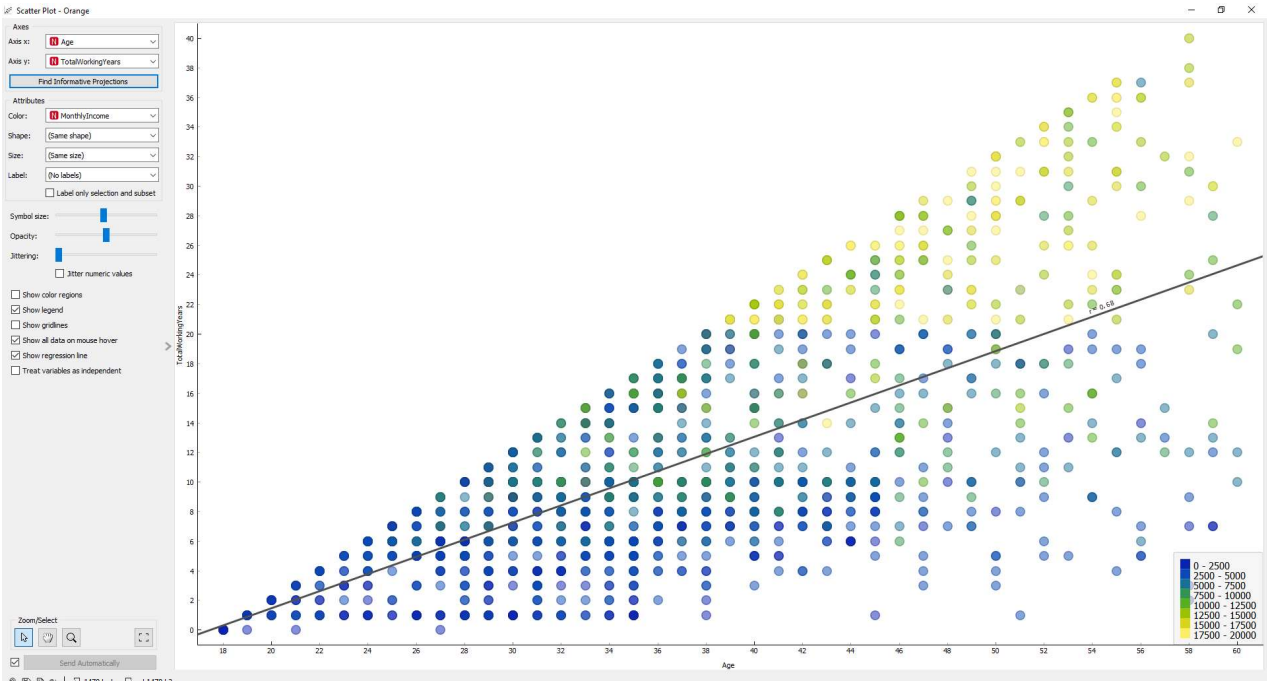
#		Univar. reg.	RReliefF
1	JobLevel	13676.938	0.299
2	MonthlyRate	1.781	0.191
3	DailyRate	0.087	0.182
4	HourlyRate	0.366	0.175
5	DistanceFromHome	0.425	0.172
6	JobSatisfaction	0.075	0.161
7	RelationshipSatisfaction	0.983	0.159
8	EnvironmentSatisfaction	0.058	0.158
9	NumCompaniesWorked	33.567	0.146
10	Education	13.358	0.140
11	Age	483.762	0.138
12	TotalWorkingYears	2177.973	0.135
13	StockOptionLevel	0.043	0.127
14	YearsWithCurrManager	197.136	0.127

In the test & score, we see the R^2 is high, which means that more than 90% of the attrition is explained with this model and based on these variables.



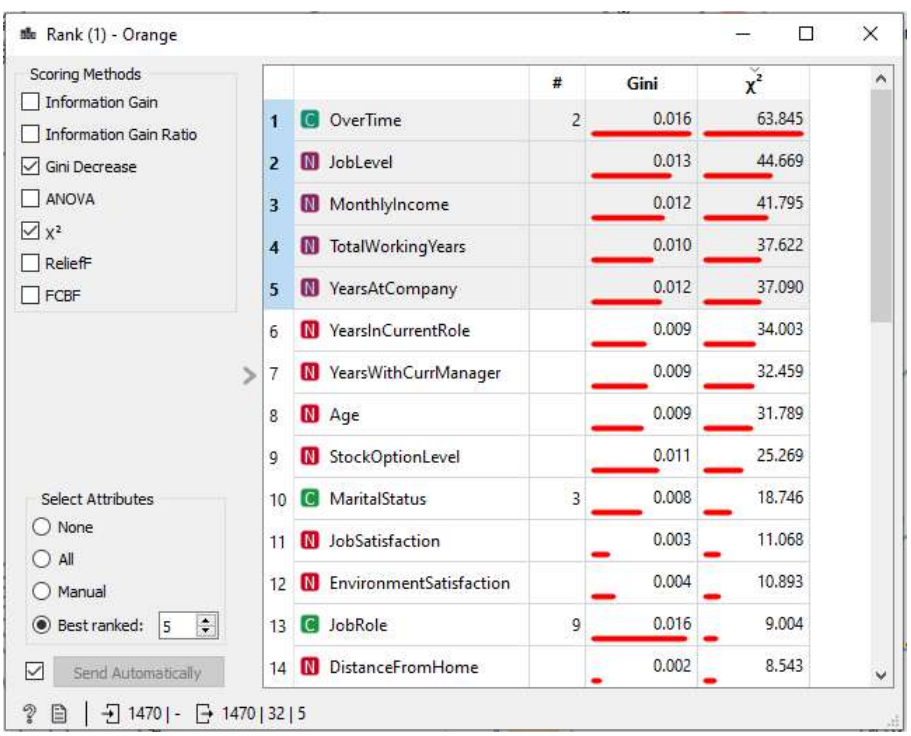
Model	MSE	RMSE	MAE	R2
Linear Regression	2120553.574	1456.212	1107.420	0.904

The following scatter plot shows a positive and relatively strong correlation between age, total working years and monthly income:



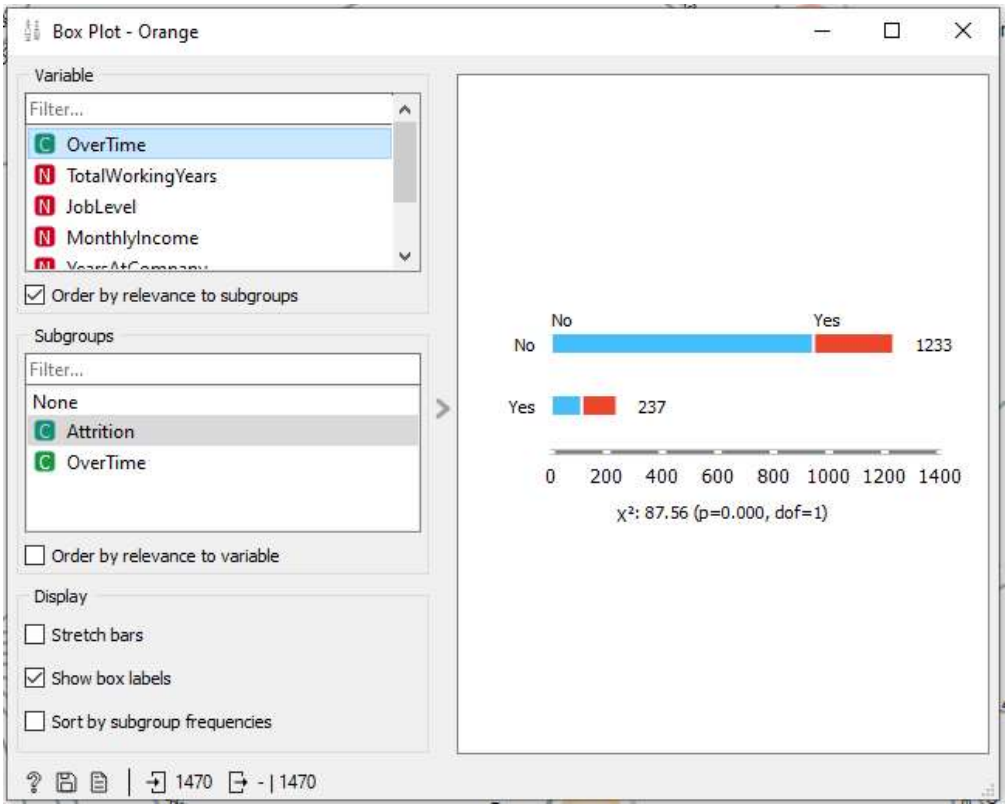
4.2. Categorical Target

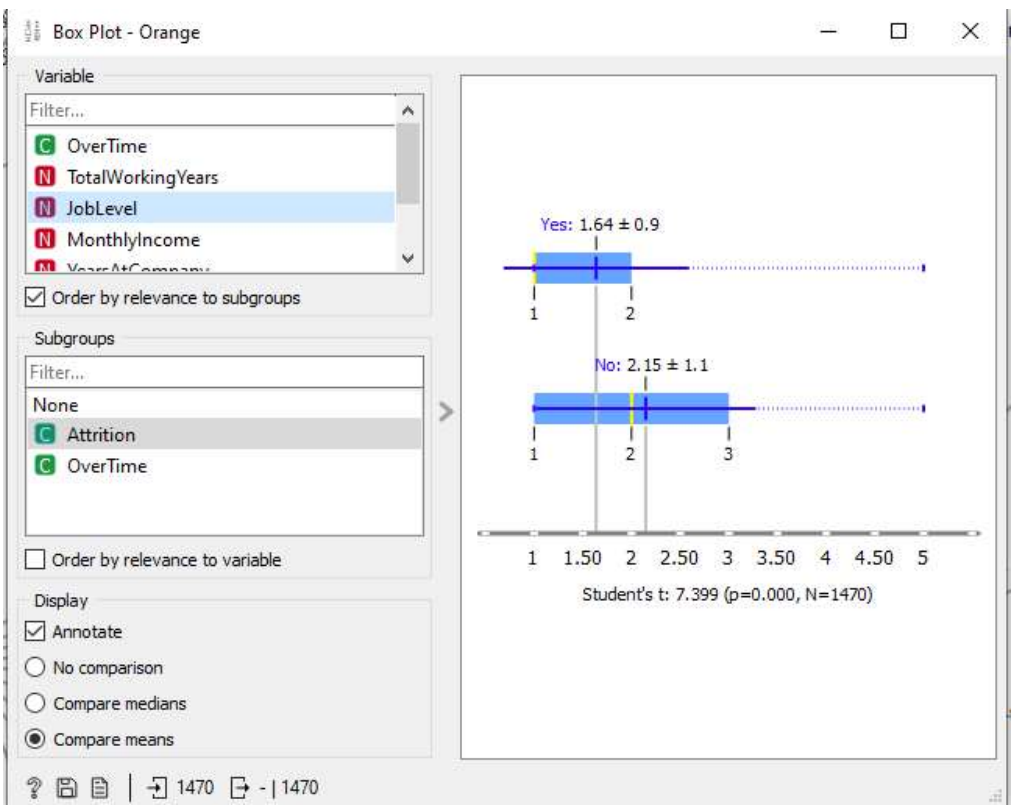
We now analyze the categorical variable attrition as the target variable. The Ranking shows the variables with higher chi-square: overtime, job level and monthly income, which consequently have a higher influence on attrition.

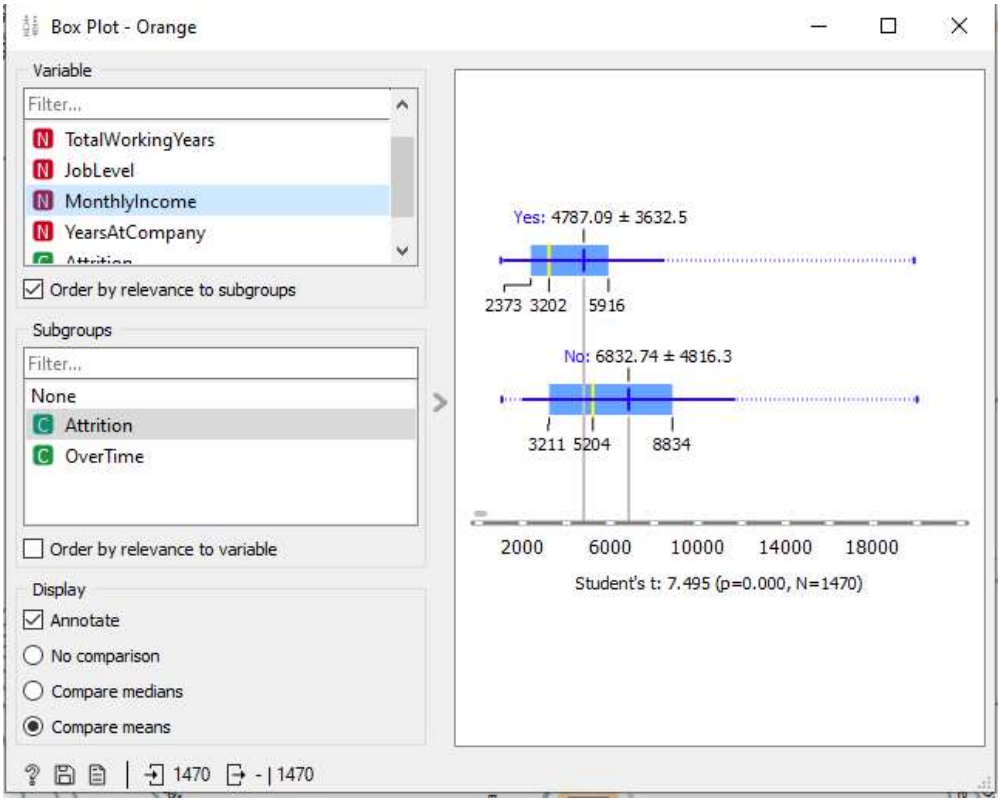


		#	Gini	χ^2
1	OverTime	2	0.016	63.845
2	JobLevel		0.013	44.669
3	MonthlyIncome		0.012	41.795
4	TotalWorkingYears		0.010	37.622
5	YearsAtCompany		0.012	37.090
6	YearsInCurrentRole		0.009	34.003
7	YearsWithCurrManager		0.009	32.459
8	Age		0.009	31.789
9	StockOptionLevel		0.011	25.269
10	MaritalStatus	3	0.008	18.746
11	JobSatisfaction		0.003	11.068
12	EnvironmentSatisfaction		0.004	10.893
13	JobRole	9	0.016	9.004
14	DistanceFromHome		0.002	8.543

The box plots shows how these variables differenciate workers which suffer attrition or not.



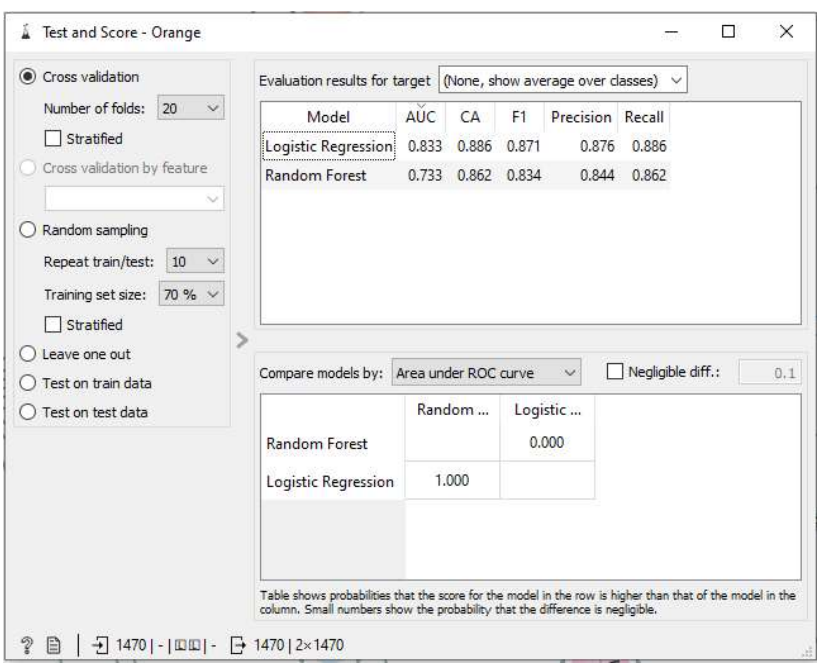




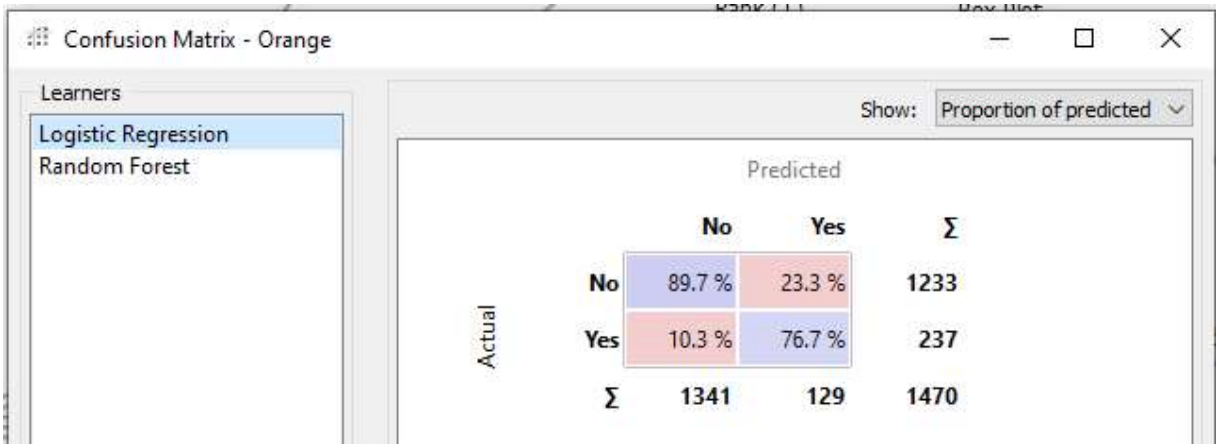
5. REGRESSION MODELS

We're addressing now the regression models for the target variable attition, where we applied 2 predictive models: logistic regression and random forest, and 1 third model for comparison, the PCA (Principal Component Analysis).

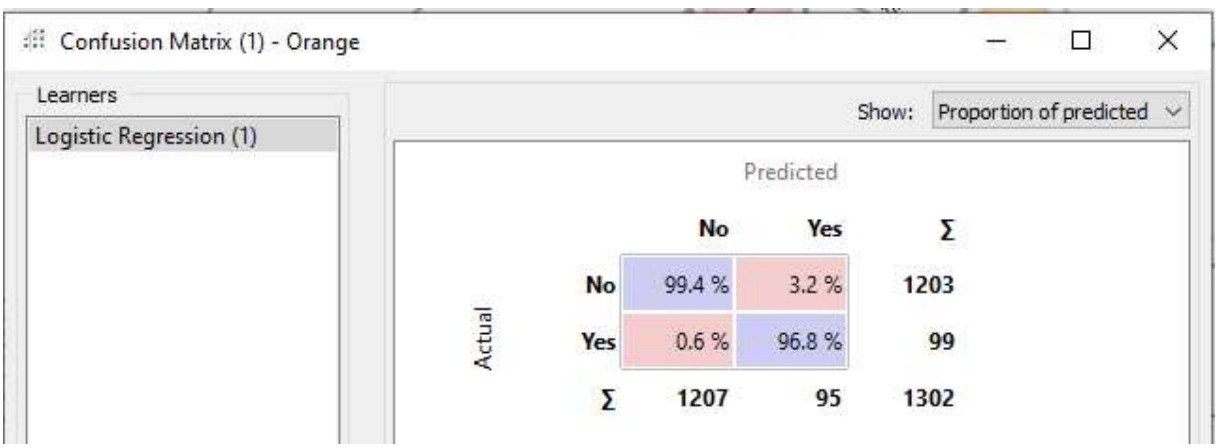
Among the first 2 Models, the logistic reaches the best score on the CA and F1:



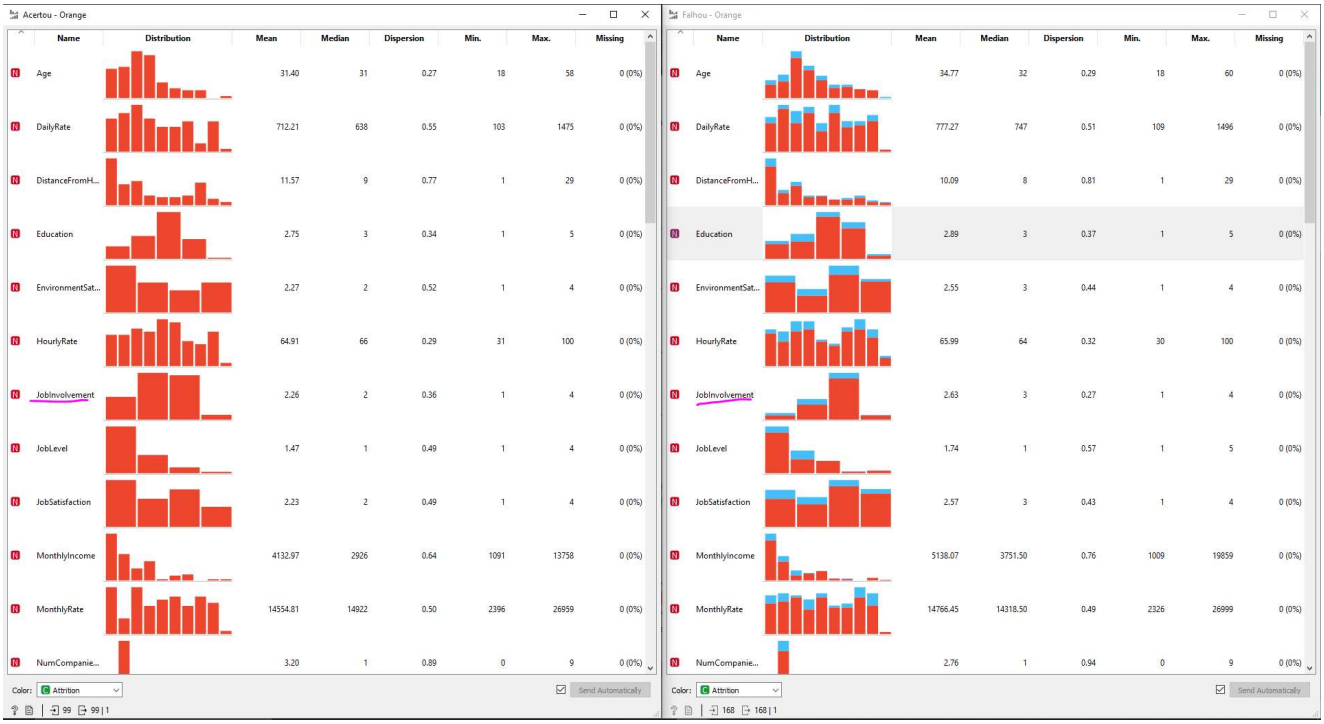
Hereunder the confusion matrix in the logistic regression:



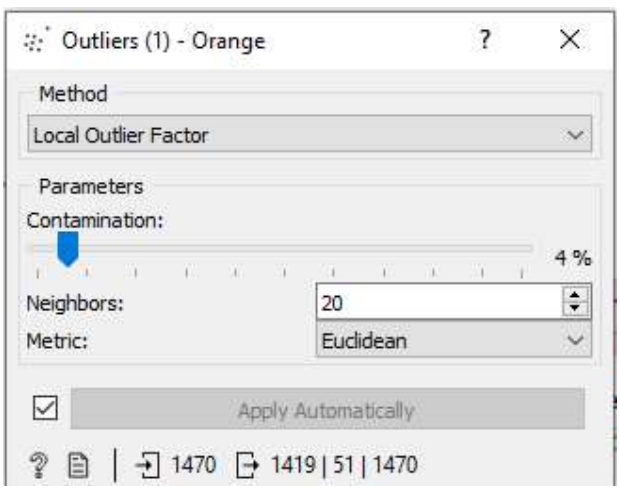
Filtering the cases in which the model is correct (Yes-Yes and No-No), we can see in the confusion matrix the model improves its correctness above 95%, as expected:

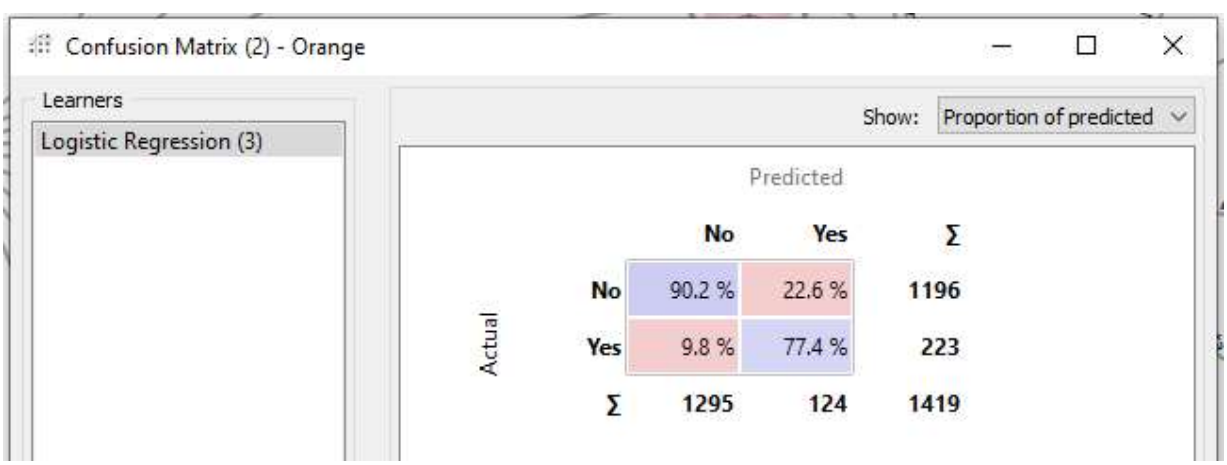


If we compare the incorrect cases (Yes-No, No-Yes), the distributions differ in some of the variables, which means those may need a deeper analysis in order to enhance the prediction model performance (e.g. Job Involvement).

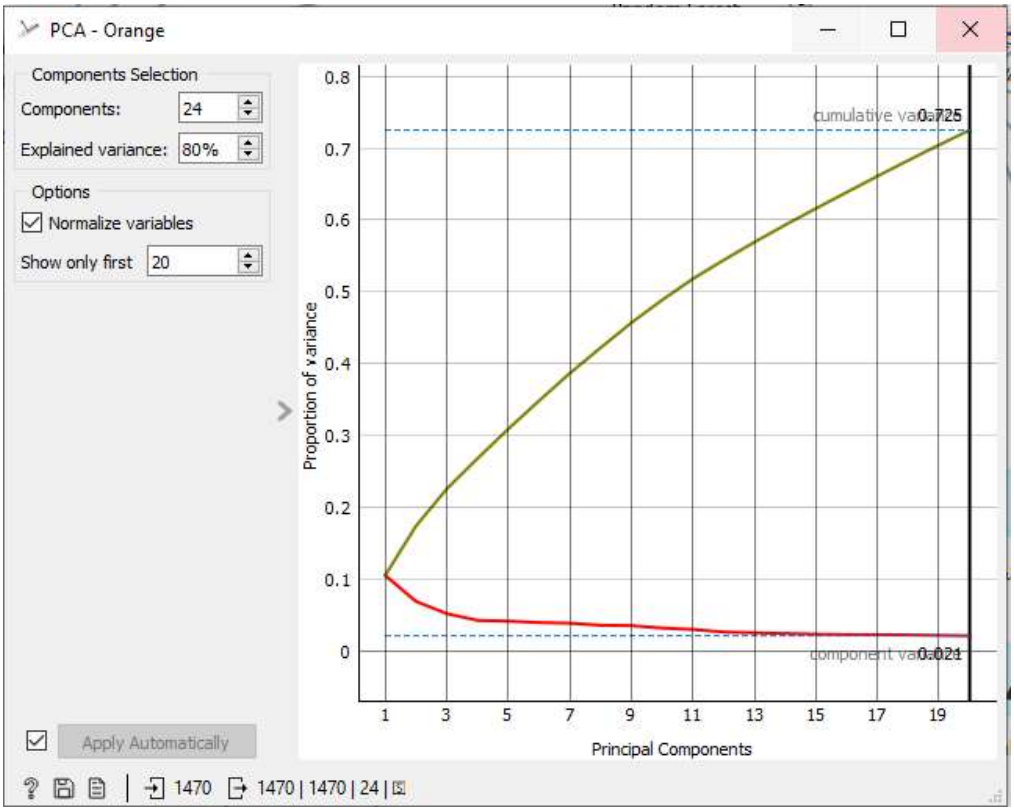


If we remove the outliers, we can see the correctness rate slightly improves:

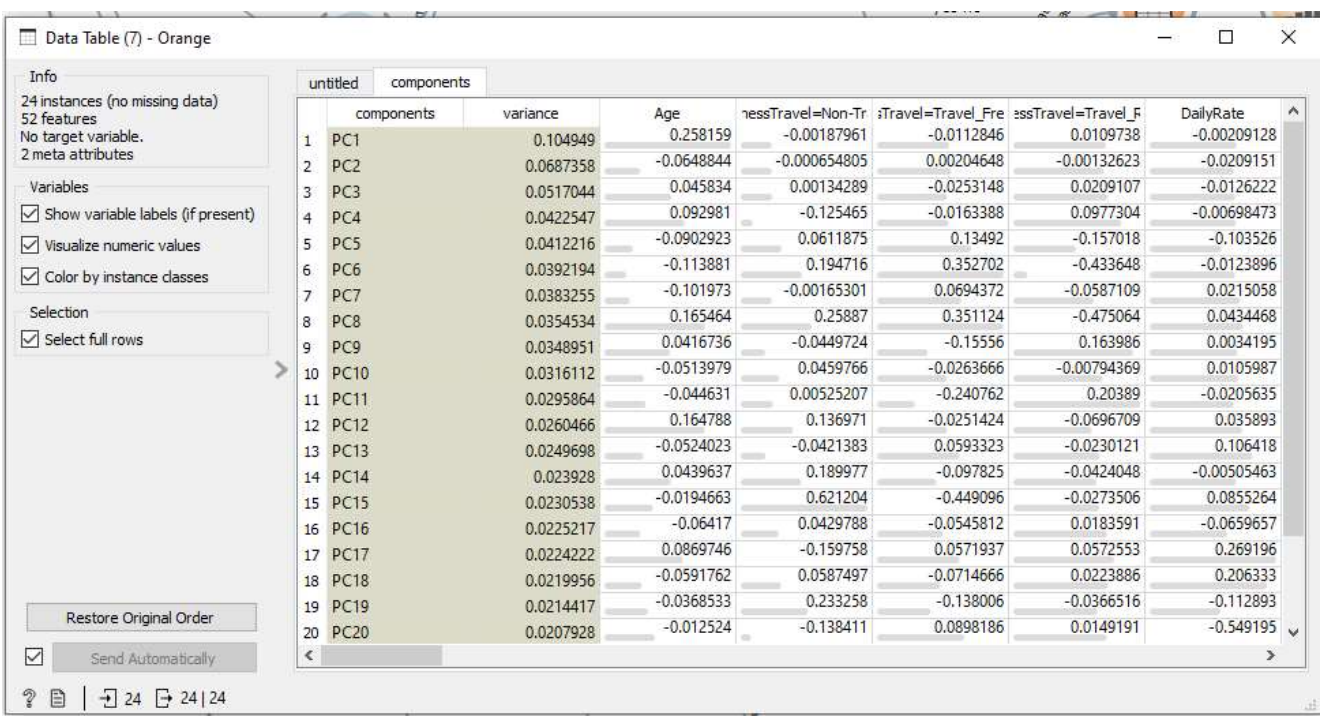




In the 3rd test applying the PCA, we were able to reduce from 33 to 24 variables, and by explaining 80% of the variance.

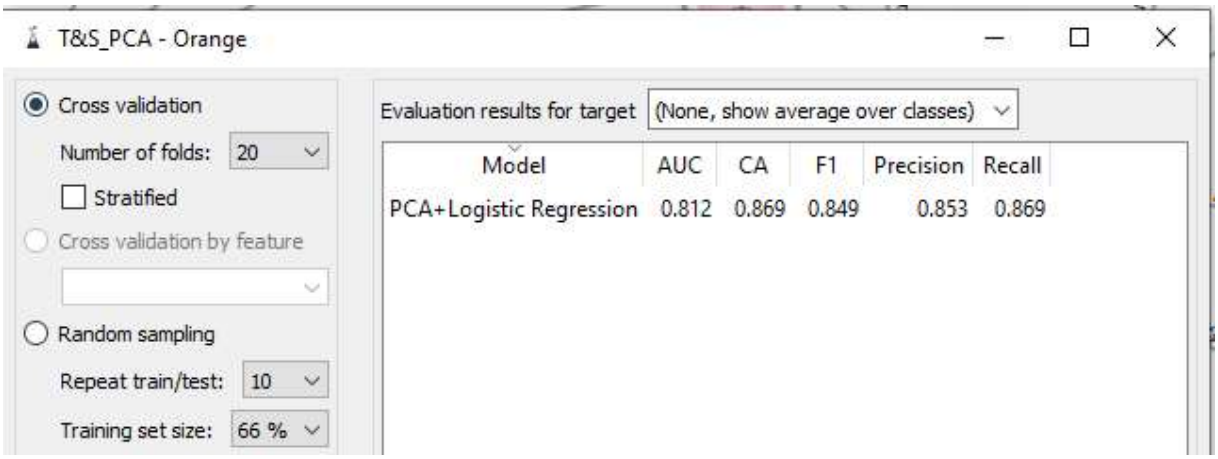


The data table shows the different weight of the components in the total variance.



	components	variance	Age	essTravel=Non-Tr	Travel=Travel_Fre	essTravel=Travel_F	DailyRate
1	PC1	0.104949	0.258159	-0.00187961	-0.0112846	0.0109738	-0.00209128
2	PC2	0.0687358	-0.0648844	-0.000654805	0.00204648	-0.00132623	-0.0209151
3	PC3	0.0517044	0.045834	0.00134289	-0.0253148	0.0209107	-0.0126222
4	PC4	0.0422547	0.092981	-0.125465	-0.0163388	0.0977304	-0.00698473
5	PC5	0.0412216	-0.0902923	0.0611875	0.13492	-0.157018	-0.103526
6	PC6	0.0392194	-0.113881	0.194716	0.352702	-0.433648	-0.0123896
7	PC7	0.0383255	-0.101973	-0.00165301	0.0694372	-0.0587109	0.0215058
8	PC8	0.0354534	0.165464	0.25887	0.351124	-0.475064	0.0434468
9	PC9	0.0348951	0.0416736	-0.0449724	-0.15556	0.163986	0.0034195
10	PC10	0.0316112	-0.0513979	0.0459766	-0.0263666	-0.00794369	0.0105987
11	PC11	0.0295864	-0.044631	0.00525207	-0.240762	0.20389	-0.0205635
12	PC12	0.0260466	0.164788	0.136971	-0.0251424	-0.0696709	0.035893
13	PC13	0.0249698	-0.0524023	-0.0421383	0.0593323	-0.0230121	0.106418
14	PC14	0.023928	0.0439637	0.189977	-0.097825	-0.0424048	-0.00505463
15	PC15	0.0230538	-0.0194663	0.621204	-0.449096	-0.0273506	0.0855264
16	PC16	0.0225217	-0.06417	0.0429788	-0.0545812	0.0183591	-0.0659657
17	PC17	0.0224222	0.0869746	-0.159758	0.0571937	0.0572553	0.269196
18	PC18	0.0219956	-0.0591762	0.0587497	-0.0714666	0.0223886	0.206333
19	PC19	0.0214417	-0.0368533	0.233258	-0.138006	-0.0366516	-0.112893
20	PC20	0.0207928	-0.012524	-0.138411	0.0898186	0.0149191	-0.549195

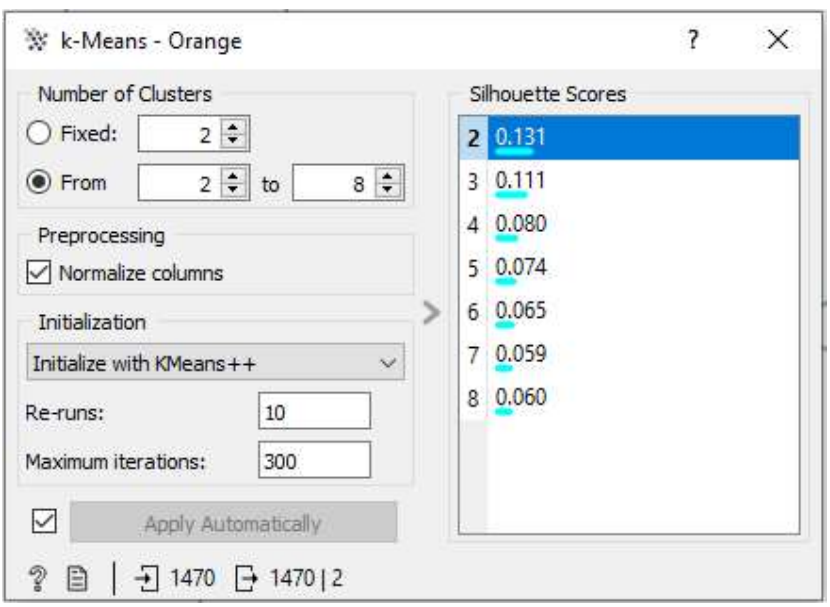
In the test & score, we see that PCA + Logistic test has a slightly lower score than Logistic test, CA = 0.869 instead of 0.884. However, this PCA improves a little over the random forest.



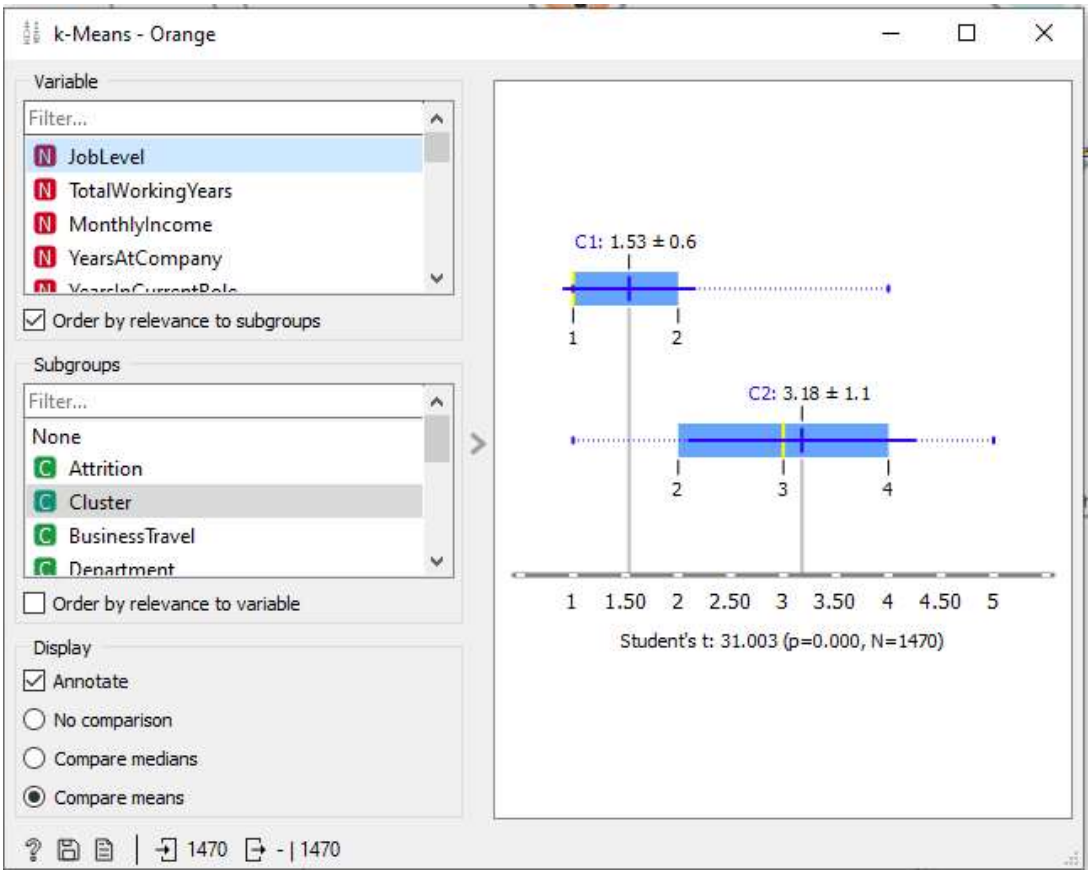
6. CLUSTERS

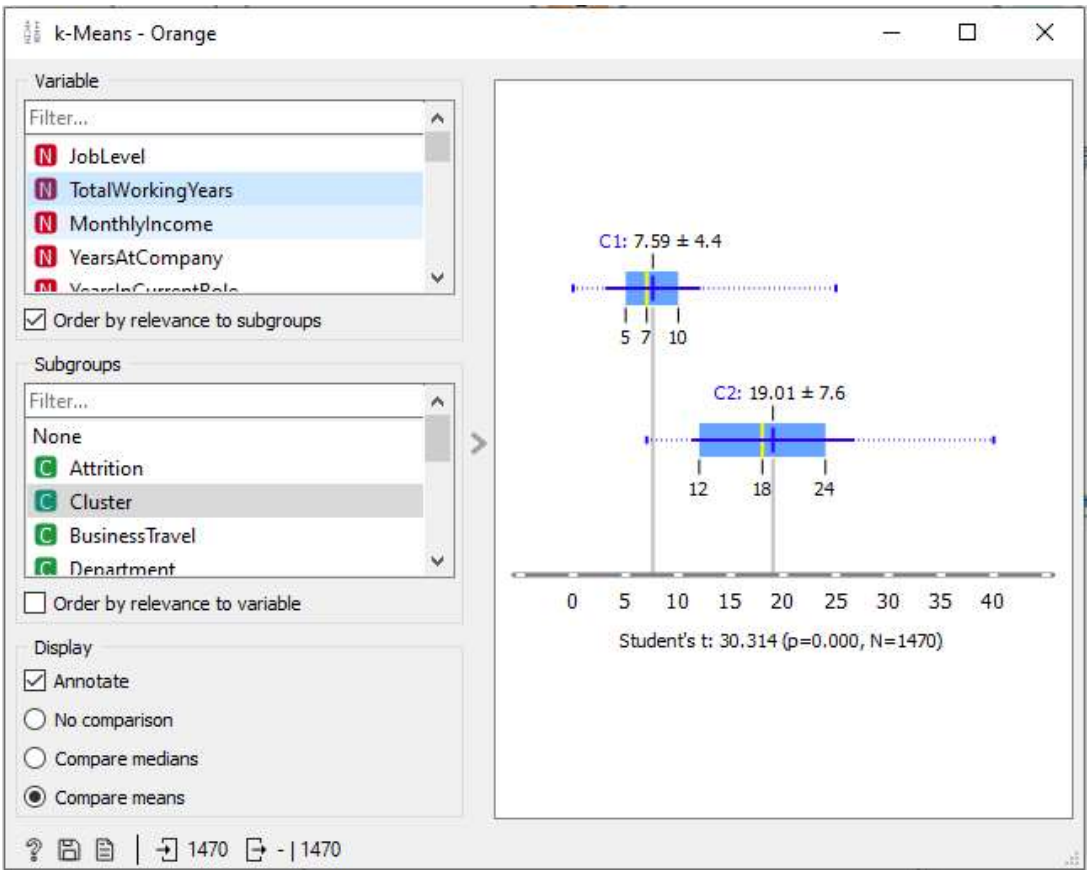
Finally, in our clusters analysis, we've used and compared 2 methods: K-Means and Hierarchical.

In the K-means method, the silhouette scores delivers a higher score for 2 clusters, which are called C1 and C2.

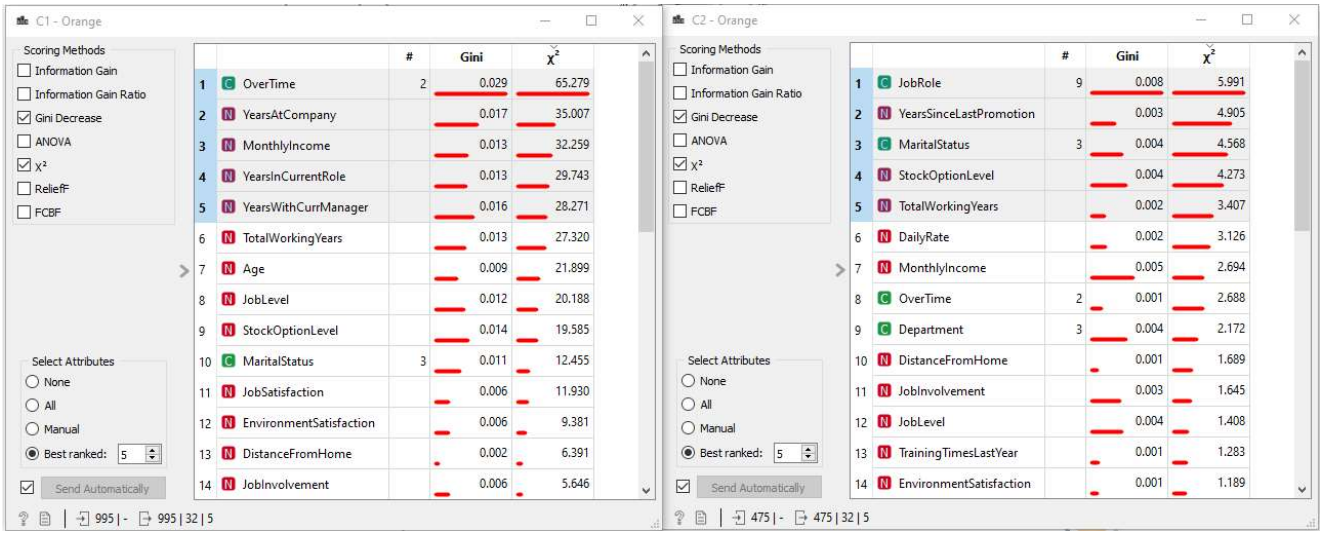


The box plot shows which variables differenciate the 2 groups: job level, total working years, monthly income, among others.


k-Means - Orange
Variable
Filter...
JobLevel
TotalWorkingYears
MonthlyIncome
YearsAtCompany
Order by relevance to subgroups
Subgroups
Filter...
None
Attrition
Cluster
BusinessTravel
Department
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
1470
- | 1470
C1: 1.53 ± 0.6
1
2
C2: 3.18 ± 1.1
2
3
4
1 1.50 2 2.50 3 3.50 4 4.50 5
Student's t: 31.003 (p=0.000, N=1470)

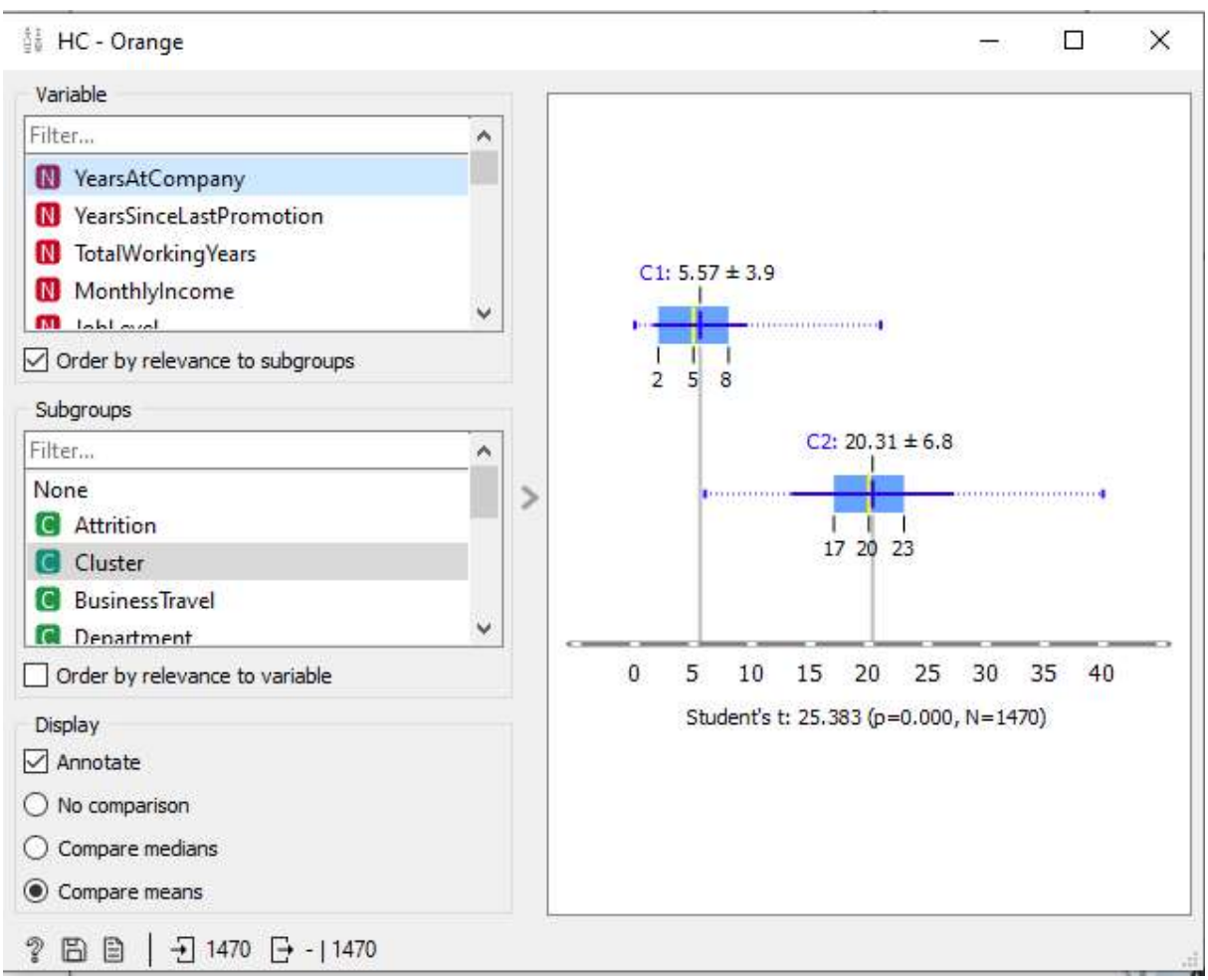

k-Means - Orange
Variable
Filter...
JobLevel
TotalWorkingYears
MonthlyIncome
YearsAtCompany
Order by relevance to subgroups
Subgroups
Filter...
None
Attrition
Cluster
BusinessTravel
Department
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
1470
- | 1470
C1: 7.59 ± 4.4
5 7 10
C2: 19.01 ± 7.6
12 18 24
0 5 10 15 20 25 30 35 40
Student's t: 30.314 (p=0.000, N=1470)

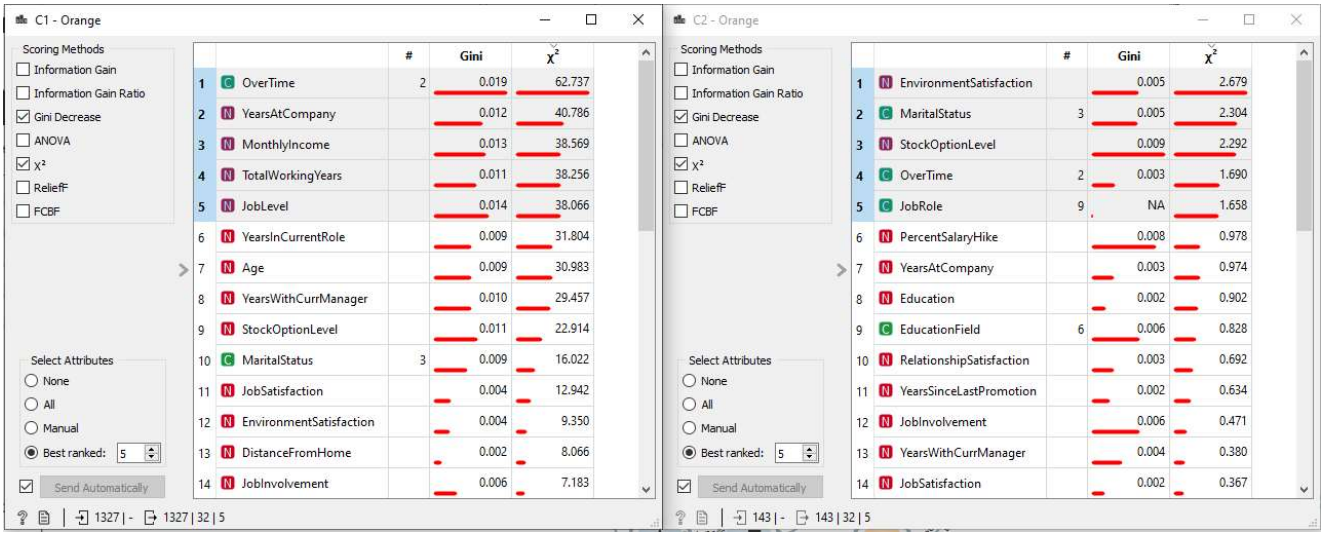
The Rank shows the variables which have a bigger weight in both groups, C1 and C2, which are also different from each other.



Hierarchical clustering

The same operation, I separated the sample into 2 groups. We can see that in this method the most differentiating variables are years at company, years since last promotion and total working years.





Conclusion

In this study we've covered statistical analysis of a dataset called attrition, namely the following topics:

Descriptive statistics:

- Most of the workers do not suffer attrition
- The average age is aprox. 37 years old.
- Job satisfaction: more people are satisfied than dissatisfied
- Overtime: the majority of workers do not work overtime

Hypothesis tests showed distribution is not normal for all variables, and there's no difference between genders in terms of salary.

Several predictive models were applied:

- A multiple linear regression model for target variable monthly income. Variables such as job level, rate and distance from home have the strongest correlation with monthly income.

- Predictive models applied for the target variable attrition: logistic regression, random forest and a third model for comparison PCA + logistic regression.

The variables with a stronger correlation with attrition are overtime, job level and monthly income.
Logistic regression reached a better performance in terms of predicting attrition.

Finally, in our clusters analysis, we used 2 methods: K-Means and Hierarchical. For each of them, we considered 2 groups: C1 and C2.
Some of the variables wich differenciate the 2 groups are job level, total working years and monthly income, among others.